UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 102 — Videregående kurs i statistikk.
Eksamensdag:	Mandag 3. juni 1991.
Tid for eksamen:	09.00 – 15.00

Oppgavesettet er på 10 sider.

Vedlegg:	Tabell over standard normalfordelingen, kji-kvadrat fordelingen og F-fordelingen.
Tillatte hjelpemidler:	Formelsamlinger for ST 101 og ST 102. Lommeregner.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

Vannverket i en kommune ønsker å undersøke forekomsten av en bestemt type bakterier i drikkevannet. For dette formålet tas vannprøver ulike steder i drikkevannsreservoiret og antall bakterier i hver av disse telles.

Vi vil anta at antall bakterier i en vannprøve på v liter er Poissonfordelt med parameter $v\lambda$. Her angir altså λ bakterietettheten pr. liter. Videre antas antall bakterier i ulike prøver å være stokastisk uavhengig.

Vannverket tar n prøver med volum $v_1, v_2, \ldots, v_n$, henholdsvis, og registrerer antall bakterier $X_1, X_2, \ldots, X_n$ i disse.

a) Vis at

$$\hat{\lambda} = \sum_{i=1}^{n} X_i \Big/ \sum_{i=1}^{n} v_i$$

(Fortsettes side 2.)

er en forventningsrett estimator for λ og bestem var($\hat{\lambda}$).

Av helsemessige grunner er det viktig at bakterietettheten ikke overstiger 10 pr. liter. Vannverket ønsker derfor å teste nullhypotesen

$$H_0 : \lambda \leq 10$$

mot alternativet

$$H_1 : \lambda > 10$$

b) Utled en test med signifikansnivå 5% for det aktuelle hypoteseprøvingsproblemet når vannverket tar 5 vannprøver hver med volum 1/2 liter. Du kan benytte at $\hat{\lambda}$ fra punkt a) vil være tilnærmet normalfordelt.

Hvilken konklusjon kan vannverket trekke hvis nullhypotesen forkastes? Hva hvis den ikke forkastes?

c) Hva blir teststyrken for testen i punkt b) for $\lambda = 15$?

Hvordan må vannverket ta vannprøvene for å oppnå en teststyrke på 0,99 for $\lambda = 15$?

Oppgave 2.

La Y_{ij} for $i = 1, \ldots, m$ og $j = 1, \ldots, k$ være uavhengige og normalfordelte stokastiske variable alle med varians σ^2 og med

$$EY_{ij} = \mu_j \qquad \text{for} \quad i = 1, \ldots, m.$$

La $\bar{Y}_{\cdot j} = \frac{1}{m}\sum_{i=1}^{m} Y_{ij}$ og $\bar{Y}_{\cdot\cdot} = \frac{1}{k}\sum_{j=1}^{k} \bar{Y}_{\cdot j}$ og sett $n = mk$.

a) La $S_0^2 = SSE/(n-k)$, hvor

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{m}(Y_{ij} - \bar{Y}_{\cdot j})^2.$$

Gi en begrunnelse for at S_0^2 er en forventningsrett estimator for σ^2.

Anta i punktene b) og c) at $x_1, \ldots, x_k$ er gitte tall slik at

$$\mu_j = a + bx_j \tag{1}$$

(Fortsettes side 3.)

for parametere a og b. Vi lar $\bar{x} = \frac{1}{k}\sum_{j=1}^{k} x_j$.

b) Vis at minste kvadraters estimatorer for a og b er

$$\hat{A} = \bar{Y}_{\cdot\cdot} - \hat{B}\bar{x}$$

og

$$\hat{B} = \frac{\sum_{j=1}^{k}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})(x_j - \bar{x})}{\sum_{j=1}^{k}(x_j - \bar{x})^2}$$

c) Gi en begrunnelse for at $S_1^2 = m \cdot SSR/(k-2)$, hvor

$$SSR = \sum_{j=1}^{k}(\bar{Y}_{\cdot j} - \hat{A} - \hat{B}x_j)^2\,,$$

er en forventningsrett estimator for σ^2 og for at S_1^2 er stokastisk uavhengig av S_0^2 gitt i punkt a).

Vi ønsker nå å teste nullhypotesen om at (1) holder, dvs. hypotesen om at μ_j-ene varierer lineært med x_j-ene.

d) Forklar hvorfor det er rimelig å forkaste nullhypotesen om at (1) holder for store verdier av testobservatoren

$$F = S_1^2/S_0^2\,,$$

hvor S_0^2 og S_1^2 er gitt i punktene a) og c). Vis at under nullhypotesen er F F-fordelt med $k-2$ og $n-k$ frihetsgrader.

Oppgave 3.

Det er viktig at sprøytemidler som benyttes i frukthager ikke er skadelig for biene og at de heller ikke har en frastøtende effekt på disse. I et forsøk ønsket en å undersøke om et bestemt sprøytemiddel ("lime sulphur") har en frastøtende effekt på bier. For dette formålet ble sukkerløsninger tilsatt ulike konsentrasjoner av sprøytemiddelet plassert i et forsøkskammer sammen med ca. 100 bier i to timer. Fem beholdere ble benyttet for hver av konsentrasjonene 0.000064%, 0.00032%, 0.0016%, 0.008%, 0.04%, 0.2% og 1%. Ved slutten av forsøket målte en hvor mye biene hadde forbrukt av sukkerløsningene (målt i milligram) for hver av de 35 beholderene.

Resultatet av forsøket ble som følger:

(Fortsettes side 4.)

Konsentrasjon i prosent	0.000064	0.00032	0.0016	0.008	0.04	0.2	1
Logaritmen (med grunntall 5) av konsentrasjonen (x_j)	−6	−5	−4	−3	−2	−1	0
	77	57	39	36	9	14	5
Forbruk	92	20	51	22	17	4	4
i mg.	24	90	55	27	16	7	2
(Y_{ij})	72	69	47	20	15	10	3
	71	71	61	51	19	4	2
Gjennomsnitt ($\bar{Y}_{\cdot j}$)	67,2	61,4	50,6	31,2	15,2	7,8	3,2

$$SSE = 6370,4 \qquad SSR = 126,17$$

Bemerk at vi for å forenkle beregningene nedenfor også har oppgitt verdiene av $SSE = \sum_j \sum_i (Y_{ij} - \bar{Y}_{\cdot j})^2$ og $SSR = \sum_j (\bar{Y}_{\cdot j} - \hat{A} - \hat{B}x_j)^2$, definert i punktene a) og c) i Oppgave 2, hvor her x_j angir logaritmen (med grunntall 5) av konsentrasjonen.

a) Tegn opp bienes gjennomsnittlige forbruk av sukkerløsningene som funksjon av logaritmen (med grunntall 5) av konsentrasjonen i et diagram. Benytt testen i Oppgave 2 punkt d) til å teste hypotesen om at det er en lineær sammenheng mellom forbruk og logaritmen av konsentrasjonen. Hva blir konklusjonen?

Det kan være aktuelt i den foreliggende situasjonen å foreta analysen basert på transformerte data av forbruket. Vi vil betrakte to transformasjoner av dataene, nemlig logaritme-transformerte data:

$$Y'_{ij} = \ln Y_{ij},$$

og logit-transformerte data:

$$Y''_{ij} = \ln(Y_{ij}/(95 - Y_{ij})).$$

De transformerte datasettene blir som følger:

(Fortsettes side 5.)

Logaritmen (med grunntall 5) av konsentrasjonen	−6	−5	−4	−3	−2	−1	0
	4.34	4.04	3.66	3.58	2.20	2.64	1.61
Logaritmen	4.52	3.00	3.93	3.09	2.83	1.39	1.39
av forbruket	3.18	4.50	4.01	3.30	2.77	1.95	0.69
(Y'_{ij})	4.28	4.23	3.85	3.00	2.71	2.30	1.10
	4.26	4.26	4.11	3.93	2.94	1.39	0.69
Gjennomsnitt	4.12	4.01	3.91	3.38	2.69	1.93	1.10

$$SSE = 5,467 \qquad SSR = 1,812$$

Logaritmen (med grunntall 5) av konsentrasjonen	−6	−5	−4	−3	−2	−1	0
	1.45	0.41	−0.36	−0.49	−2.26	−1.76	−2.89
Logit-	3.42	−1.32	0.15	−1.20	−1.52	−3.13	−3.13
transformert	−1.09	2.89	0.32	−0.92	−1.60	−2.53	−3.84
forbruk	1.14	0.98	−0.02	−1.32	−1.67	−2.14	−3.42
(Y''_{ij})	1.09	1.09	0.58	0.15	−1.39	−3.13	−3.84
Gjennomsnitt	1.20	0.81	0.13	−0.76	−1.69	−2.54	−3.42

$$SSE = 23,948 \qquad SSR = 0,9755$$

b) Undersøk på samme måte som i punkt a) om det er lineær sammenheng mellom hver av de transformerte datasettene over bienes forbruk av sukkerløsningene og logaritmen av konsentrasjonen av sprøytemiddel.

Uansett resultatet av undersøkelsene i punktene a) og b) er det foretatt lineære regresjonsanalyser av sammenhengen mellom bienes forbruk av sukkerløsningene og logaritmen (med grunntall 5) til konsentrasjonen av sprøytemiddelet. Det er foretatt separate analyser for de utransformerte dataene (Y_{ij}-ene) så vel som for de to transformerte datasettene (Y'_{ij}-ene og Y''_{ij}-ene).

Disse analysene ga følgende estimater for parameterene a og b (gitt i (1) i Oppgave 2) og verdier for R^2:

(Fortsettes side 6.)

Utransformerte data (Y_{ij})

Parameter	Estimat	Standardfeil
a	−2.05	4.44
b	−11.96	1.231
$R^2 = 0.74$		

Logaritme-transformerte data (Y'_{ij})

Parameter	Estimat	Standardfeil
a	1.47	0.157
b	−0.515	0.043
$R^2 = 0.81$		

Logit-transformerte data (Y''_{ij})

Parameter	Estimat	Standardfeil
a	−3.29	0.265
b	−0.800	0.073
$R^2 = 0.78$		

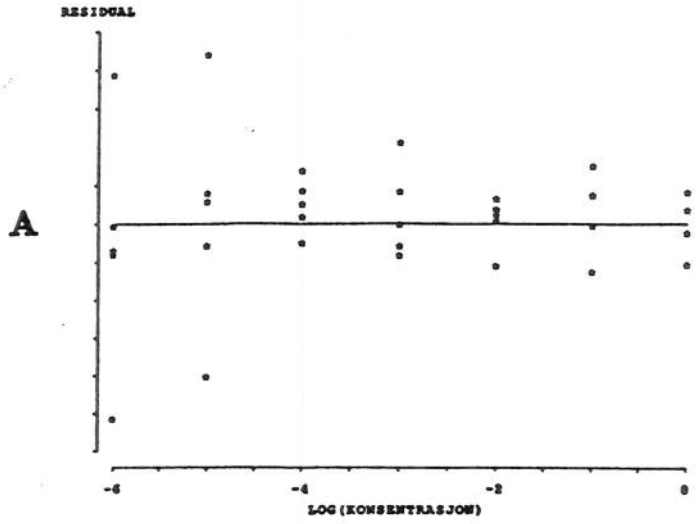
Videre er det lagd plott av residualene i disse regresjonsanalysene mot logaritmen (med grunntall 5) til konsentrasjonen av sprøytemiddel. Disse plottene er gitt på neste side, men ikke (nødvendigvis) i samme rekkefølge som resultatene av regresjonsanalysene ble presentert. (Skalaen langs y-aksen er med vilje fjernet fra plottene.)

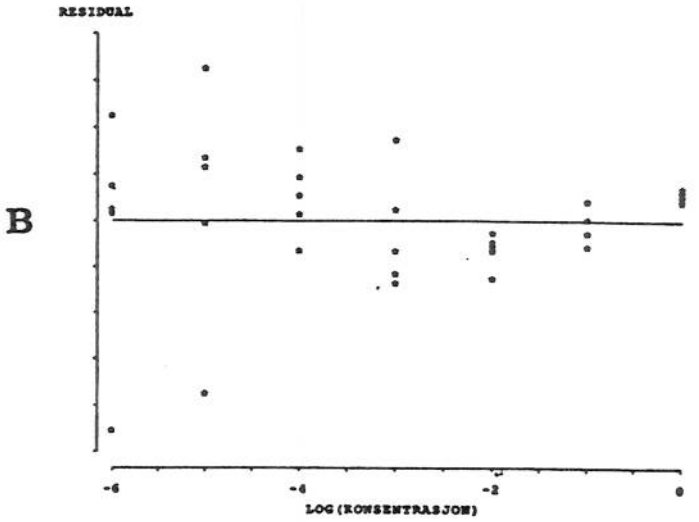
c) Identifiser hvilket residualplott som hører til hver av de tre regresjonsanalysene som er foretatt. Svaret skal begrunnes.

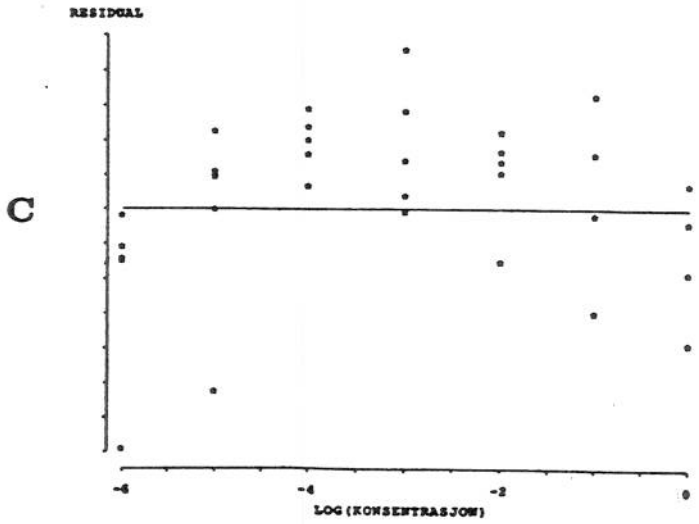
d) Diskuter for hver av de tre regresjonsanalysene om forutsetningene for å benytte lineær regresjon ser ut til å være oppfylt. Kommenter også resultatene fra testingen i punktene a) og b) i lys av denne diskusjonen.

Hvilken av regresjonsanalysene (om noen) syns du gir en tilfredsstillende analyse av dataene.

(Fortsettes side 7.)







(Fortsettes side 8.)

Oppgave 4.

I denne oppgaven skal vi studere og illustrere en teknikk som benyttes i biologisk forskning for å finne ut om en populasjon av dyr eller planter er i "genetisk likevekt" eller om populasjonen er utsatt for genetiske endringer (for eksempel på grunn av innvandring fra en populasjon med annen genetisk sammensetning).

Vi konsentrerer oss om et bestemt genpar som kan bestå av to ulike gener som vi kaller A og a. De ulike genkombinasjonene for dette genparet blir følgelig AA, Aa og aa. Vi forutsetter at vi for et gitt individ kan identifisere hvilken av de tre genkombinasjonene individet har (enten fordi de tre genkombinasjonene gir individer som ser ulike ut eller ved hjelp av genteknologiske teknikker).

Vi antar at den populasjonen som studeres er (uendelig) stor, og betrakter et tilfeldig uttrukket individ. Sannsynlighetene for at dette individet (uansett kjønn) skal ha genkombinasjonene AA, Aa og aa, henholdsvis, antas å være p_{AA}, p_{Aa} og p_{aa}.

Et individ i neste generasjon arver tilfeldig ett av farens gener og ett av morens gener. Sannsynligheten for at individet skal arve genet A fra sin far er følgelig

$$\theta = p_{AA} + \frac{1}{2}p_{Aa}\,.$$

Dette er også sannsynligheten for at individet skal arve A fra sin mor. Vi antar videre at han- og hunindividene ikke har noen spesielle preferanser i partnervalget, slik at pardannelsen foregår tilfeldig. Dette betyr at et avkom får genene fra sin far og sin mor uavhengig av hverandre.

a) Vis at sannsynlighetene for at et individ i neste generasjon skal ha genkombinasjonene AA, Aa og aa blir

$$p_{AA} = \theta^2, \quad p_{Aa} = 2\theta(1-\theta) \quad \text{og} \quad p_{aa} = (1-\theta)^2\,, \tag{2}$$

henholdsvis. (En kan lett innse at disse sannsynlighetene vil være uendret også i de derpå følgende generasjoner.)

En populasjon som tilfredsstiller (2) er i "genetisk likevekt", såkalt Hardy-Weinberger likevekt, og avvik fra denne likevekten betyr at populasjonen er i endring f.eks. på grunn av innvandring.

(Fortsettes side 9.)

b) Betrakt en populasjon som er i Hardy-Weinberger likevekt, dvs. som tilfredsstiller (2). Vi trekker et tilfeldig utvalg på n individer fra populasjonen og registrerer deres genkombinasjoner. La X_{AA}, X_{Aa} og X_{aa} angi antallet av de n som har genkombinasjonene AA, Aa og aa, henholdsvis. Sett opp likelihood funksjonen og vis at sannsynlighetsmaksimeringsestimatoren for θ blir

$$\hat{\theta} = \frac{2X_{AA} + X_{Aa}}{2n}$$

c) Forklar hvordan du vil teste hypotesen om at populasjonen er i Hardy-Weinberger likevekt, dvs. hypotesen om at (2) holder.

Vi vil benytte overstående til å studere om torskebestanden nær Danmark og i Østersjøen er i Hardy-Weinberger likevekt. Torsk er blitt fanget ved Lolland, Bornholm og Åland og forekomsten av torsk med ulike typer hemoglobin har blitt studert. Vi nøyer oss med å kalle hemoglobintypene for AA, Aa og aa. En fant følgende fordeling av genkombinasjonene for de ulike stedene.

	Genkombinasjon			
Sted	AA	Aa	aa	Totalt
Lolland	27	30	12	69
Bornholm	14	20	52	86
Åland	0	5	75	80

d) Test hypotesen om at torskepopulasjonen ved Lolland er i Hardy-Weinberger likevekt. Test også hypotesen om at torskebestanden ved Bornholm er i Hardy-Weinberger likevekt. Kan du foreta en test om at populasjonen ved Åland er i Hardy-Weinberger likevekt basert på de foreliggende dataene? Kommenter resultatene.

Da Bornholm ligger mellom Lolland og Åland kan en tenke seg at torsken fanget ved Bornholm består av en blanding av en østlig (Åland) og en vestlig (Lolland) torskestamme som hver for seg er i Hardy-Weinberger likevekt. I så fall blir sannsynlighetene for genkombinasjonene AA, Aa og aa for torsk fanget ved Bornholm

$$\begin{aligned} p_{AA,B} &= \alpha\theta_{\text{Å}}^2 + (1-\alpha)\theta_L^2 \\ p_{Aa,B} &= 2\alpha\theta_{\text{Å}}(1-\theta_{\text{Å}}) + 2(1-\alpha)\theta_L(1-\theta_L) \\ p_{aa,B} &= \alpha(1-\theta_{\text{Å}})^2 + (1-\alpha)(1-\theta_L)^2 \end{aligned} \tag{3}$$

(Fortsettes side 10.)

hvor $\theta_{\text{Å}}$ og θ_L er θ-verdiene som inngår i (2) for torskestammene ved Åland og Lolland, henholdsvis, og α er en ukjent parameter som angir blandingsforholdet.

La $\hat{\theta}_{\text{Å}}$ være estimatoren gitt i punkt b) beregnet på grunnlag av tallene fra Åland, og la $\hat{\theta}_L$ være den tilsvarende estimatoren basert på dataene fra Lolland.

e) Forklar hvorfor en rimelig estimator for α kan fastlegges ved ligningen

$$\hat{\alpha}\hat{\theta}_{\text{Å}} + (1-\hat{\alpha})\hat{\theta}_L = \frac{2X_{AA,B} + X_{Aa,B}}{2n_B}$$

hvor $X_{AA,B}$ og $X_{Aa,B}$ er antall torsk med genkombinasjon AA og Aa fanget ved Bornholm og n_B er totalt antall torsk fanget der. Beregn $\hat{\alpha}$ på grunnlag av de observerte data.

f) Vurder, uten å foreta noen formell hypotesetesting, om "blandingsmodellen" (3) virker rimelig.

SLUTT